Missing data exploration in air quality data set using R-package data visualisation tools

Shamihah Muhammad Ghazali, Norshahida Shaadan, Zainura Idrus
Faculty of Computer and Mathematical Sciences, University of Technology MARA Shah Alam, Malaysia

Article Info

Article history:

Received Oct 16, 2019
Revised Dec 28, 2019
Accepted Jan 16, 2020

Keywords:

Air quality
Missing data
Missing data mechanism
Missing data pattern
Visualisation

ABSTRACT

Missing values often occur in many data sets of various research areas. This has been recognized as data quality problem because missing values could affect the performance of analysis results. To overcome the problem, the incomplete data set needs to be treated using imputation method. Thus, exploring missing values pattern must be conducted beforehand to determine a suitable method. This paper discusses on the application of data visualisation as a smart technique for missing data exploration aiming to increase understanding on missing data behaviour which include missing data mechanism (MCAR, MAR and MNAR), distribution pattern of missingness in terms of percentage as well as the gap size. This paper presents the application of several data visualisation tools from five R-packages such as visdat, VIM, ggplot2, Amelia and UpSetR for data missingness exploration. For an illustration, based on an air quality data set in Malaysia, several graphics were produced to illustrate the contribution of the visualisation tools in providing insight on the pattern of missingness. Based on the results, it is shown that missing values in air quality data set of the chosen sites in Malaysia behave as missing at random (MAR) with small percentage and long gap sizes of missingness.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Shamihah Muhammad Ghazali,
Faculty of Computer and Mathematical Sciences,
University of Technology MARA,
40450 Shah Alam, Selangor Darul Ehsan, Malaysia.
Email: shamihah.ghazali@gmail.com

1. INTRODUCTION

Nowadays, most cities in the country have their own monitoring stations to control the environmental quality, which measures five different parameters in the air namely Sulphur Dioxide (SO2), Nitrogen Oxides (NO2), Carbon Monoxide (CO), Ozone (O3) and PM10 of Air Pollutant Index (API) [1], mainly pollutants problem are affected by the fossil fuel, especially NO2 that lead to climate change and acid rain [2]. In air quality studies, it is imperative to understand the temporal variation of the air quality data from different stations [3], data characteristics behind different backgrounds of environment and a huge amount of data recorded by hourly for every day [4]. Additionally, air quality data are often impacted by the existence of missing values in the data set, which then caused a big problem in most studies[5, 6].

Missing data is a frequent problem occurred in research fields including environmental studies of air quality studies, industry field, in surveys, census data as well as clinical research [7-11]. Missing data is a big challenge to researchers to make a good decision in their study and to understand the data. Missing data in air quality dataset often occurs from failure in the measurements units, machine malfunctions during some bad seasonal weather, computer system crashes, human errors and staying off-line for several days [12]. In some cases, the missing values are the most important points that interpret or hold important information about the other points where important awareness and the missing values must be replaced by a meaningful value that imputed using imputation methods. However, some of the commonly used imputation methods are based only on the available points and deleting the whole columns or rows that results in missing values, which can yield biased estimates of the parameters and can lose a lot of important information from the deleted points. The most important characteristics to be considered when choosing the most appropriate method for imputation are mechanism of missingness and the pattern of missingness [8, 9, 13]. Appropriate visualisation tools may help researchers in gaining more knowledge about the data set and the pattern of missingness in the dataset.

In the context of statistical analysis, the data analysts face a big challenge when dealing with a high dimensional and large dataset, which may lead to problems like a noise, error and over fitting [14]. Thus, a visualisation tool for a big data can give a better understanding about the data information and provide an attractive way of graphical and pictorial presentation to identify the relationships as well as pattern, trends, distribution, missing data and outliers in groups of data set [15] using an implemented plot. Visualisation of missing data helpful in the exploration and identification of the missing data, pattern of missingness, mechanism of missingness, as well as the relationship of missing and available values and the gap size of missingness in the data set. Visualisation is one of the important tools for data analytics in providing valuable information for exploring missing data [16] through colour, position and shape that help the researchers in seeing pattern and anomaly in the data [17]. The exploration of data visualisation on the missing data can help in investigating the pattern accurately without introducing misleading patterns and masking data properties. By using a data visualisation method in presenting the data, one can prevent poor handling of missing and wrong interpretation of the data [18], the visualisation of pattern of missingness can help in improving the quality of the data, with respect to the assessing data quality control on the data with a high error of invalid and missing data [19]. Therefore, the aim of this paper is to graphically present the pattern as well as the gap size of missing data and to find whether the data is missing at random or not in the air quality data set for Malaysia. For this paper, Section 2 discusses the research methodology that emphasizes on three visualisation components and tools for exploring missing data which include missing values mechanism, distribution of missingness (percentage) and the sequence of missingness (gap size) pattern. Section 3 highlights the result of analysis and finally, Section 4 provides the study summary and conclusion.

2. RESEARCH METHOD

2.1. The air quality data

In this study, several illustration of the visualisation application were conducted using Air quality data which was obtained from the Department of Environment Malaysia involving Klang, Shah Alam and Petaling Jaya air quality monitoring sites. The data set consists of daily by hourly PM10 level recorded from year 2011-2017. The data set is arranged in data frame or matrix consisting of 24 columns (hours) and 2557 rows (days) for each station.

2.2. Data Visualisation and framework

There are various visualisation methods available for analysing and exploring missing values in R-packages. In this study, the exploration of missing data pattern using visualisation tools is divided into three components of steps with the aims Step 1: Detecting the mechanism of missing values; Step 2: Exploring the percentage of missing values distribution and Step 3: Evaluating the gap size of missingness pattern. In short, conceptual framework of the analysis procedure involved is shown in Figure 1.

2.2.1. Detection of the missing values mechanism

There are three types of missing data mechanism described by [3, 20, 21] known as MCAR (missing completely at random), MAR (missing at random) and MNAR (missing not at random). These mechanisms are the reason why the data is missed and explained on the pattern, types, relationship between the available values and missing values of the variable in the data matrix. MCAR is most commonly used in the imputation assumption, which is defined as missingness as it is independent from other missing and observed values. Meanwhile, MAR is less strict because the missingness is allowed to be dependent on the observed values, but does not depend on the other missing values. Generally, the missingness pattern is supposed to be MCAR except the data is proven to be missing dependent on the observed values and have a strong evidence to support it. If the data is rejected to be MCAR, it is supposed to be MAR, unless there is an evidence to support that the data is not MAR, then the data has to be assumed as MNAR. Thus, visualisation of the missing data determines the relationship of the missingness of variables. There are many visualisations to be used in displaying the mechanism of missing data between variables. For example, scatter plot, matrix plot, two parallel coordinates plot, ggplot and correlation matrix plot. The function used for detecting the missingness mechanism in this paper was vis_cor() from Visdat package [22].

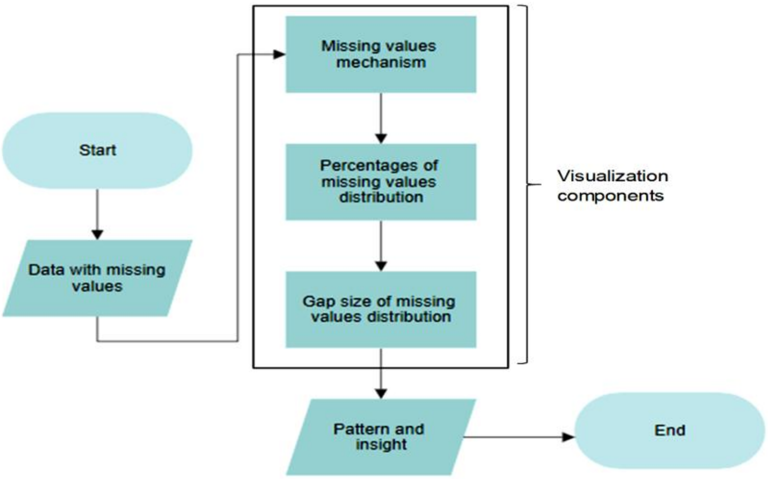

Figure 1. Conceptual framework of the visualisation components procedure

2.2.2. Exploring the percentages of missing values distribution

To explore the pattern of missingness percentage in a data set, a good starting point is to examine the number of missingness in every variable in order to get a quick summary on missingness in the data set with a graphical summary. The function of gg_miss_var() from the ggplot2 package [23] can provide a good graphical summary. The summary presents the frequency of the missing values in every variable. The pattern of missingness can be also visualised by plotting a map of the whole dataset that provides a wide overview on the location of missing data. This missingness map also visualise the whole data frame at once, providing information of whether or not the data is missing in every variable, percentage of missingness for the whole data frame and amount of missingness in each column or variable. The functions that can be used for plotting the missingness maps are missmap() from the Amelia package [24], vis_miss() and vis_data from the Visdat package [22] and heatmap() from VIM package [25].

2.2.3. Evaluating on the gap size of missing values distribution

Gaps size or empty space in the dataset shows the data in the data matrix that are not applicable, which is known as NA's. A dataset often contains missing data where from a very short gap up to a very long gap of missingness, the sizes of gap are determined by the period of time the data are missing from the datasets. However, the short or large gap in the dataset may bring a noise to the dataset. In order to investigate the gaps size of missingness in the dataset, visualisation approach from some of functions in R packages to be used. The packages have the ability to evaluate the gap size of missingness in the dataset included the aggregate plots from VIM packages may help in evaluating the gaps size as these plots can measure the number of missing values contained in a single variable besides analysing certain combinations of variable with a high number of missing and available values [26], thus allows the gap size in the dataset to be measured from the combination of variables. Next, an upset plot from the UpSetR package can be used to visualise the pattern of missingness and the combinations of missingness across the cases for every variable in the dataset. Thus, the function can be used for visualising the gap sizes of missingness are aggr() from VIM package [25], gg_miss_case from ggplot2 package [23] and gg_miss_upset() from UpSetR package [27].

3. RESULTS AND DISCUSSION

3.1. Classification of missing data mechanism

Generally, the mechanism of missing air quality data is known as MAR when the missing data in air quality is occurring randomly [3, 8, 9] in which the missing value of the variable is independent on the missing value itself but dependent on the observed values. For proving a MAR pattern, an illustration of visualisation approach were applied to the air quality dataset in Klang to detect the mechanism of missing values. To determine the missingness pattern that are occurring randomly in the dataset, the correlation between variables in the dataset and the dependencies between the missing values as well as available values of the variables need were checked. At this moment, only vis_cor() function from the R package of Visdat was used to visualise the correlations between the missing values and observed values in the variables for the air quality data for the Klang, Shah Alam and Petaling Jaya stations. The visualisation correlation plot in Figure 2 shows the correlation among variables in the Klang air quality data as a heatmap. The dark red colour represents that the value of correlation coefficient between variables was 1, whereas the bright red represents that the correlation coefficient between the variables was 0.5. There were some evidences showing that the missing data was MAR as the correlation plot shows a dark red among the correlations, thus showing a high correlation between each variable nearly to the value of 1. Similarly, the MAR mechanism works when the relationship of missing data in a variable is related to other variables in the dataset [28]. To sum up, this plot shows that there is MAR pattern in the dataset. Most of the statistical software and statistical method is required MAR assumption on the data set such as multiple imputation [29].

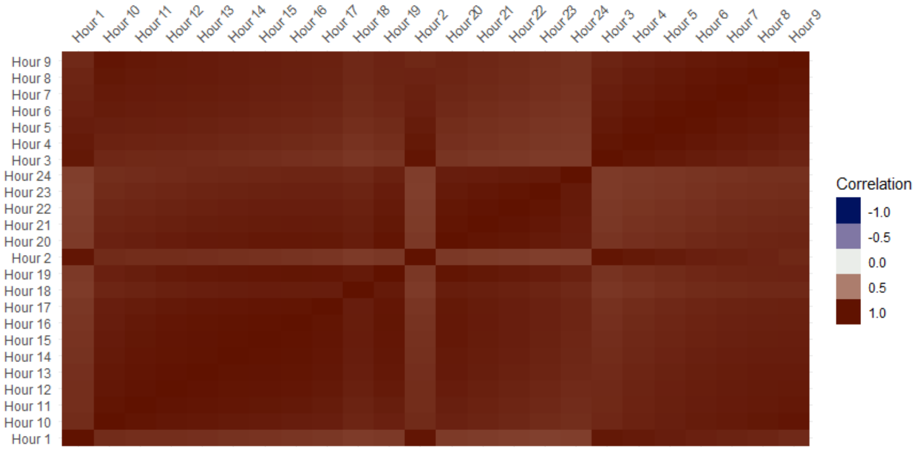

Figure 2. Correlation visualisation of nonmissing and missing values by variables in Klang station

3.2. Visual presentation of percentages of missing values distribution with R packages

There are many visualisation tools in R packages to be used for visualising the percentages of missing data pattern. The percentages of missingness for every station were determined using these functions in the R packages, there were heatmap from Visdat package, missingness maps from Visdat package and Amelia package and plot summary from ggplpt2 package. The heatmap above shows the variables containing missing values and the percentages of missing values in the air quality data for Klang data. From Figure 3, there was 8.1% of missing values and 91.9% of present values in the Klang station. These heatmaps show that the dataset has an almost complete information on all the 24 variables, but there were also observations that are missing all the time during 24 variables (24 hours). This heatmap was produced using vis_miss() from Visdat package which provides pattern and percentage summaries of missingness overall in the legend, and for each variables or column. It was apparent that all variables have a missing value with a very small proportion of missing values to the dataset. The grey colour indicates available values and appeared some structures to their missingness, which can also be used as a benchmark or referral data. Similar visualisations for map of missing values are available in other R packages such as heatmap() from VIM packages, missmap() from Amelia packages and vis_dat() from Visdat package.

The map of missingness shows the locations of missingness that occurred in the dataset, which also displays the percentages of missing and observed values for the whole dataset. The plot in Figure 4 draws a missingness maps in Shah Alam dataset using image function. The columns of the missingness maps were ordered left to the right by putting the variables with the greatest number of missingness to least number of missingness, the columns here represent the variables namely Hour 1 to Hour 24. The rows were the number of observations from bottom to the top, the rows here represent the days of the daily recorded PM10 concentration. The red colour represents the missing values with 3% of missingness and blue colour represents the observed values with 97% of availability. Figure 4 also visualises the whole set of data by looking at the missingness map to directly identify the missingness in the middle of the dataset, which are unobserved values. Additionally, the missingness map are good in illustrating how the listwise deletion will affect the statistical analysis when simply removing the missing values and a benchmark data set can be clearly seen from the plot, which can help to assume which data to be used for benchmarking purpose. This missingness map was done using missmap() from Amelia package to visualise the missingness pattern in the air quality for Shah Alam data.

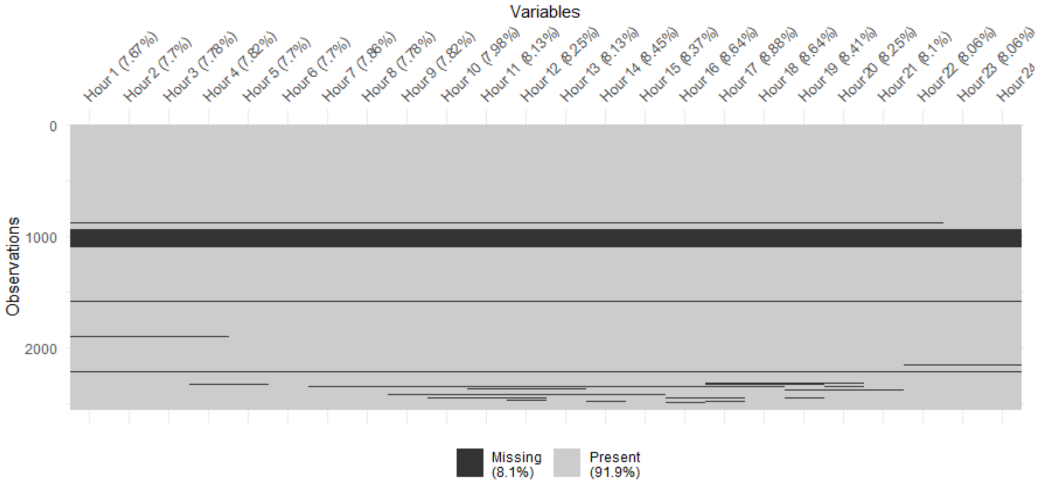

Figure 3. Heatmap visualisations of missing data for the air quality data in Klang

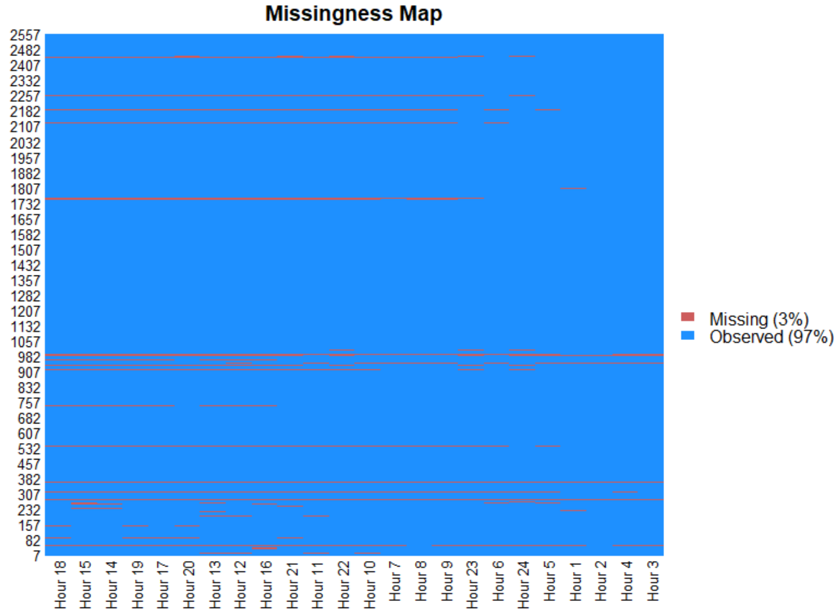

Figure 4. Missingness map visualisation of missing data for the air quality data in Shah Alam

The graphical summary shows the number of missing values in each variable visualised using gg_miss_var() from ggplot2 package. The plot in Figure 5 displays a summary on the number of missingness per variable and the rows ordered from top to bottom by arranging the variables with the greatest number of missingness to least number of missingness. From the plot, the variable of Hour 16 has been missing 75 missingness values compared to others variables, while Hour 3 recorded the lowest number of missing values with 56 missingness. From this visualisation, it is easy to identify which variable has the greatest amount of missingness and which variable has the least amount of missingness, thus making it easy to read the information of missingness in the dataset. These plots were produced using gg_miss_var() found in the naniar package developed by [30] and based on ggplot2 package to the air quality data from Petaling

Jaya station. In the broader view, the visualisation of the pattern of percentages missing values distribution would help in assessing data quality levels that can benefit the organizations in identify data errors that need to be fixed and assess whether the data in their human error or technically systems is fit to serve its intended purpose.

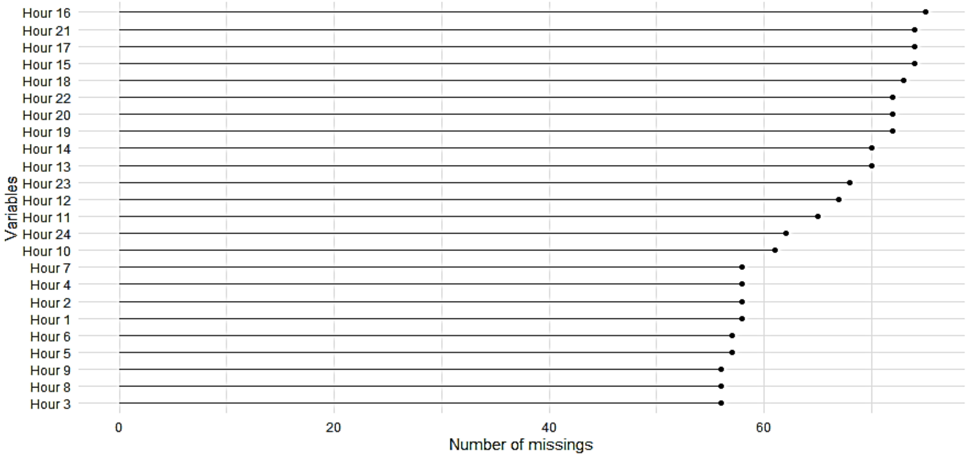

Figure 5. Graphical summary visualisation of missingness in each variable for air quality data in Petaling Jaya

3.3. Exploring the gap sizes of missingness with R packages

Gap sizes of missingness in a dataset should also be explored because the capability of different imputation methods available might depend on gap size handling. Visualisation tools using the function aggr() of VIM package, upset plot from UpSetR package and miss_case_summary() from ggplot2 package would help in exploring the gap size pattern exist. The aggregation plots shown in Figure 6, clearly displays the number of missing values contained in a single variable and the combination of variables from high to low number of missing values for Klang site data set. The plot on the left-hand side displays a bar plot for each variable while the bar height represents the proportion of missing values for each variable. For example, the highest number of missing values was in Hour 17 variable with 227 number of missing values. The plot on the right-hand side showed an aggregation plot revealing the combinations of missing values in red colour and observed values in blue colour. Additionally, the horizontal axis in the aggregation plot representing the frequencies of different combinations of variable were visualised by a small bar plot on the right side while the vertical axis represents the observations with a missing and observed values of the variables. This plot revealed the top row representing a combination where the first 20 variables of an observation have a missing value while the remaining variables have no missing values; this means that the gap sizes of missingness for this observation were 20 gaps of missingness.

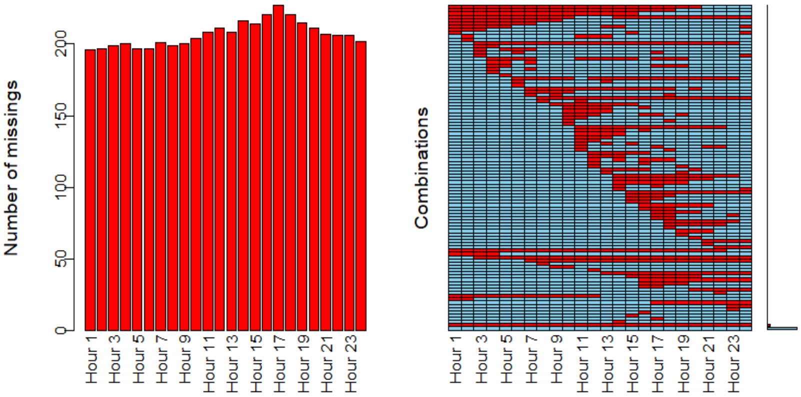

Figure 6. Aggregation plot visualisation on the proportion of missing values in a variable and combination of missing and non-missing values in the observation of air quality data for Klang station

The upset plot is a visualisation tool for giving a clearer view of missingness pattern and gap of missingness, which was achieved using the UpSetR package for the air quality data in Shah Alam. From Figure 7, the upset plot visualised the number of times certain variables are missing together with the other variables, which are the combination of missingness that shows a pattern of missingness gap in the dataset. For example, this plot displayed 12 combinations of missing values in 24 variables; the first bar showed a combination where the first 24 variables or hours of an observation have missing values in which this combination appeared 16 times in the data. Thus, it means that there were 24 gaps of missingness in the dataset that appeared for 16 times. Next example was the last bar showing a combination of 18 variables of an observation with missing values that appeared 3 times in the data, which is equivalent to 18 gaps of missingness; all the 18 variables out of 24 variables were involved except for variable 1 to variable 6. The information of the variables missing can be seen through the black dot in the bar plot below.

The size of missing gaps is shown in the graphical summary of missingness in each case; this plot was visualised using gg_miss_case function from ggplot2 package and shown using the air quality dataset from Petaling Jaya stations. There were missing values in every 24 variables of the data shown in graphical summary of missingness (Figure 5), forming a plot summary for each case in the variables that are missing. From plot in Figure 8, there were between 1 to 24 missing value in cases, which means that the largest size of missing gap was 24 and the smallest size of missing gap was 1. There were many missing values in the data with majority of the missingness were missing with 24 gaps of missingness, which were found with 38 cases and involved 24 variables. Additionally, there were 16 cases with 1 gap of missingness in the dataset. The information from the graphical summary of missnginess cases would help in identifying the gap sizes of missingness in the data.

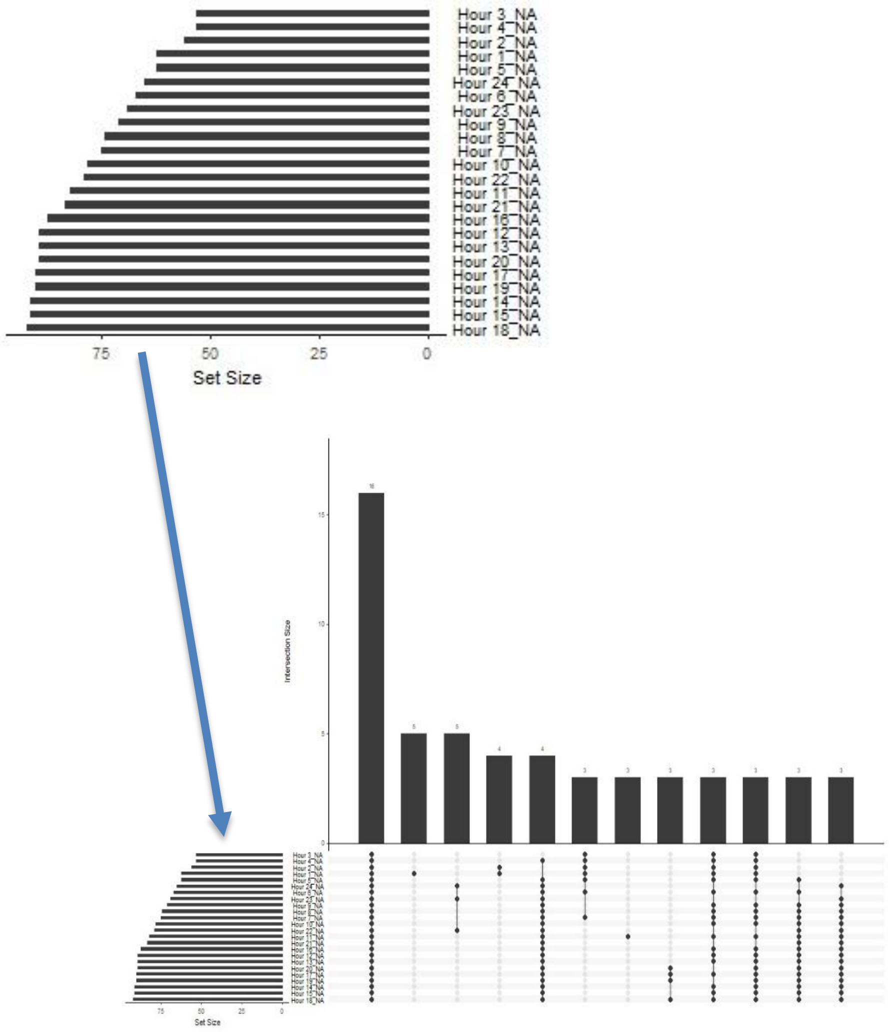

Figure 7. Upset plot visualisation on the missingness pattern and combination sets of missing variables in the air quality data for Shah Alam station

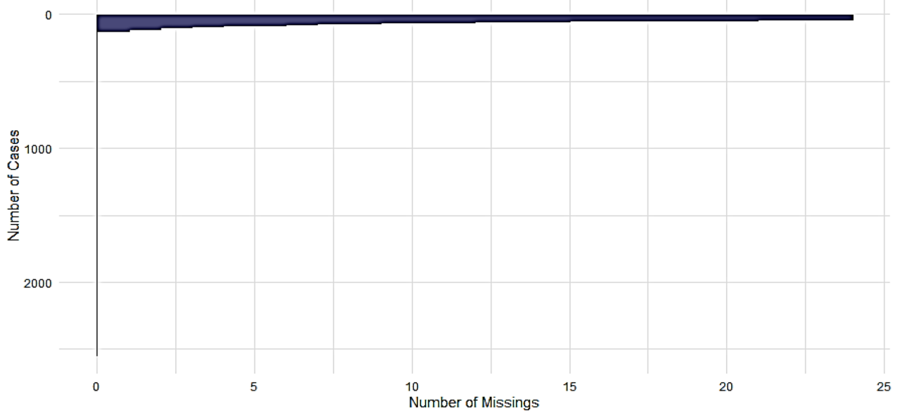

Figure 8. Graphical summary visualisation of missingness in each case for air quality data in Petaling Jaya

4. CONCLUSION

This paper has described graphical methods for visualising and exploring the missing data as a modern approach to increase understanding on missing data pattern. The visualisation of missing data can help in supporting the analysis decision and in clearly understanding the details about the missing data structures and patterns. Additionally, visualisation is also one of statistical requirements for a data to be tested with a graphically and adequacy checking on data quality like a missingness, outliers or skewed data distributions before further analysis testing. The visualisation work in this paper was motivated by recent development and implementation in R packages to proceed several solutions in handling and visualising the missing values in a data. The visualisation tools were from the R packages namely visdat, VIM, ggplot2, Amelia and UpSetR. These tools provide interactive graphics in visualising the missing data pattern and allow the combination of information of missingness in one variable to the other in the dataset. Linking the plot with the other function plots for a better explanation on the missing data structure, these tools provided plots and visual graphics that are very useful in visualising the missing data mechanism, missing data pattern and gaps of missingness. Implemented tools in this study could be used for other fields of dataset whether circular data, census data or environmental related data. All the mentioned packages are available on the comprehensive R archive network (CRAN). In conclusion, the implementation of these visualisation tools has achieved the study objectives to assist in detecting missing data mechanism, exploring distribution pattern of percentage missingness and evaluating the pattern and existence of gap size of missingness in air quality data set. The results of the analysis have shown that air quality data set of the chosen sites is MAR, has minor percentage of missingness and do contain large gap size. For future investigation, a comparative study on issues pertaining to the level of missingness such as minor, moderate or major is suggested.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support by the Dr. Norshahida Shaadan and Dr. Zainura Idrus from University Technology MARA Shah Alam for providing the information and guidance. The authors also would like to thank the Department of Environment Malaysia (DOE) for providing the information and data. This work is supported by the Institute of Quality and Knowledge Advancement (InQKA).

REFERENCES

[1] Department of Environment Malaysia, "Air Quality." [Online]. Available: https://www.doe.gov.my/portalv1/en/info-umum/kuality-udara/114.

[2] P. Ilamathi, V. Selladurai and K. Balamurugan, "Predictive modelling and optimization of power plant nitrogen oxides emission," in *International Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 11-18, March 2012.

[3] S. Hirabayashi and C. N. Kroll, "Single imputation method of missing air quality data for i-Tree Eco analyses in the conterminous United States," pp. 1-24, 2017.

[4] P. Chen, "Visualization of real-time monitoring datagraphic of urban environmental quality," in *EURASIPJournal on Image and Video Processing*, vol. 6, no. 42, 2019.

[5] Z. Zhang, "Big-data Clinical Trial Column Missing data imputation: focusing on single imputation," in *Hemodial. International Journal Translational Medicine*, vol. 4, no. 1, p. 9, Jan 2016.

[6] A. Mair and A. Fares, "Comparison of rainfall interpolation methods in a mountainous region of a tropical Island," in *Journal of Hydrologic Engineering*, vol. 16, no. 4, pp. 371-383, April 2011.
[7] Y. Xia, P. Fabian, A. Stohl and M. Winterhalter, "Forest climatology: Estimation of missing values for Bavaria, Germany," in *Agricultural and Forest Meteorology*, vol. 96, no. 1-3, pp. 131-144, August 1999.
[8] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," in *Atmospheric Environment.*, vol. 38, no. 18, pp. 2895-2907, Jun 2004.
[9] A. Plaia and A. L. Bondì, "Single imputation method of missing values in environmental pollution data sets," in *Atmospheric Environment*, vol. 40, no. 38, pp. 7316-7330, Dec 2006.
[10] N. M. Noor, S. Y. Ahmad, R. Nor Azam and A. Mohd Mustafa Al Bakri, "Estimation of missing values in air pollution data using single imputation techniques," in *ScienceAsia*, vol. 34, no. 3, pp. 341-345, 2008.
[11] M. N. Norazian Ramli, A. S. Yahaya, N. A. Ramli, N. F. F. M. Yusof and M. M. A. Abdullah, "Roles of imputation methods for filling the missing values: A review," in *Advanced in Environmental Biolology*, vol. 7, pp. 3861-3869, Oct 2013.
[12] N. A. Zainuri, A. A. Jemain and N. Muda, "A comparison of various imputation methods for missing values in airquality data," Sains Malaysiana, vol. 44, no. 3, pp. 449-456, 2015, doi: 10.17576/jsm-2015-4403-17.
[13] N. M. Noor and N. A. Zakaria, "Imputation methods for filling missing data in urban air pollution data for Malaysia," in *Urbanism. Arhitectură. Construcţii*, vol. 9, no. 2, pp. 159-166, Jan 2016.
[14] F. R. Kamala, P. R. J. Thangaiah and A. Info, "An improved hybrid feature selection method for huge dimensional datasets," in *International Journal of Artifical Intelligence*, vol. 8, no. 1, pp. 77-86, 2019.
[15] A. E. Shadare and C. Akujuobi, "Data visualization," in *International Journal of Engineering Research and Advanced Technology*, vol. 2, no. 12, pp. 11-16, Dec 2016.
[16] S. J. Fernstad and R. C. Glen, "Visual analysis of missing data — To see what isn't there," *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Paris, 2014, pp. 249-250.
[17] C. Eaton, C. Plaisant, and T. Drizd, "Visualizing missing data : Classification and empirical study," pp. 1-12, 2005.
[18] X. Cheng, "Interactive visualization for missing values , time series , and areal data," in *Graduate Theses and Dissertations*, 2015.
[19] L. E. Aik, T. W. Hong and A. K. Junoh, "Distance weighted K-Means algorithm for center selection in training radial basis function networks," in *International Journal of Artifical Intelligence*, vol. 8, no. 1, pp. 54-62, Mar 2019.
[20] R. J. A. Little and D. B. Rubin, "Statistical analysis with missing data, 2nd edition," in Wiley Series in Probability and Statistics, New York, Sep 2002.
[21] M. Templ and A. Alfons, "Exploring incomplete data using visualization techniques," in *Advances in Data Analysis and Clasification*, vol. 6, no. 1, pp. 29-47, April 2012.
[22] N. Tierney, "visdat: Visualising whole data frames," in *Journal of Open Source Software*, vol. 2, no. 16, pp. 355, 2017.
[23] H. Wickham, "ggplot2: Elegant graphics for data analysis," in *Springer-Verlag New York*, 2009.
[24] J. H. and G. K. and M. Blackwell, "Amelia II: A program for missing data," in *Journal of Statistical Software*, vol. 45, no. 7, pp. 1-47, Dec 2011.
[25] A. K. and M. Templ, "Imputation with the R Package VIM," in *Journal of Statistical Software*, vol. 74, no. 7, pp. 1-16, Oct 2016.
[26] Z. Zhang, "Missing data exploration : highlighting graphical presentation of missing pattern," in *Annals of Translational Madicine*, vol. 3, no. 22, pp. 356, Dec 2015.
[27] J. Conway, A. Lex and N. Gehlenborg, "UpSetR: An R package for the visualization of intersecting sets and their properties," in *Bioinformatics*, vol. 33, pp. 2938-2940, Sep 2017.
[28] S. Nakagawa, "Missing data: mechanisms, methods, and messages," in Ecological Statistics: Contemporary theory and application, 2015.
[29] E. M. Mikkelsen, D. Cronin-fenton, N. R. Kristensen and L. Pedersen, "Missing data and multiple imputation in clinical epidemiological research," in *Clinical Epidemiology*, pp. 157-166, Mar 2017.
[30] N. Tierney, B. Statistics, D. Cook and B. Statistics, "Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations," in *Monash Econometrics and Business Statistics Working Papers*, pp. 1-41, 2018.